Big Data
final project
Natural Disaster & GPDs

Jonas Brandenburg (if20b243)
Violeta Garcia Espin (if20b091)
Rawan Mousa (if20b129)

Project idea:

Obtaining Data:

Natural Disasters

- Simple Download (csv)
- 16000 recorded disasters
- Location, Date / Duration, Severity

Country GDP's

- Worldbank API
- API call per country
- Yearly GDP
- 14000 Data points
- Country, Year, GDP value

Storing Data:

Database

- CouchDB
- Implemented with Docker
- Using json files
- Scalable

Upload

- API call -> parse to object -> upload to db
- No local saving required
- Table for Natural Disasters
- Table for GDP’s

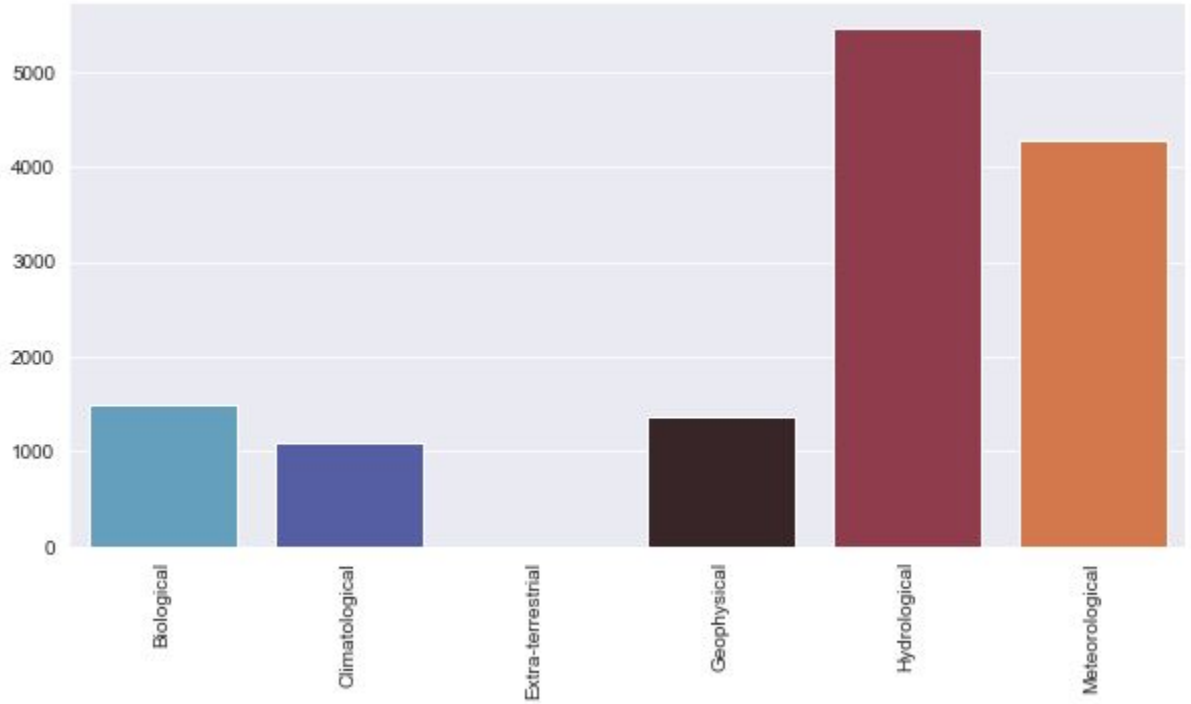
Querying data: MapReduce

```
design = { 'views': {
               'get_all_countries': { # view1 with map function
                   'map': """
                         function(doc) {
                             emit(doc.iso_country, 1)
                         }""",
                   'reduce': '_sum',
                },
                'by_total_affected': {
                    'map': """
                        function(doc) {
                            emit(doc.total_affected, doc)
                        }
                    """
                },
```

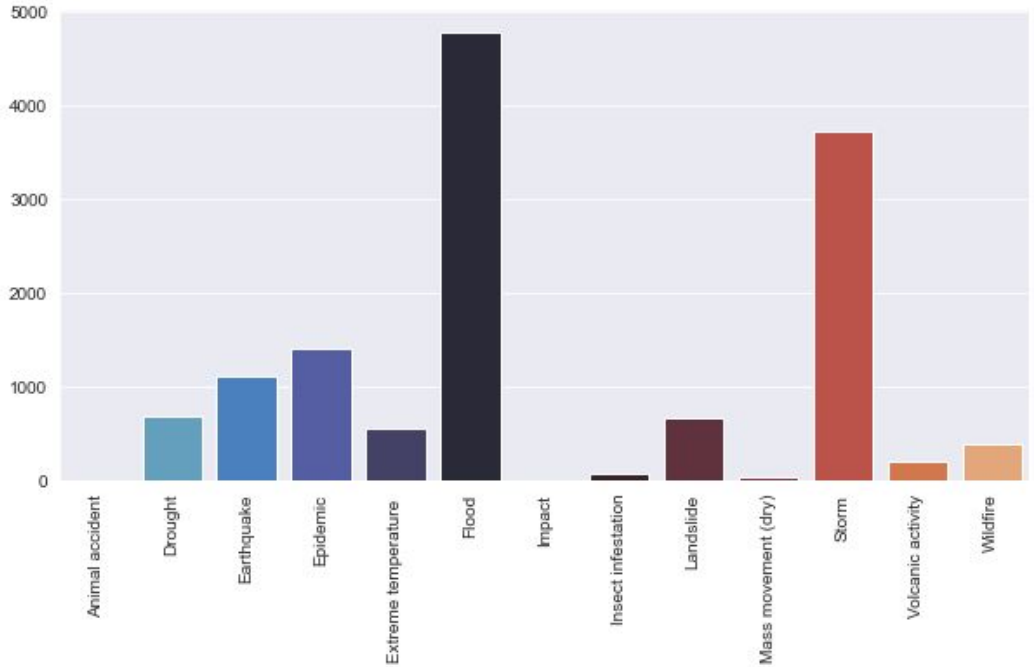
```
from mrjob.job import MRJob

class MRWordCount(MRJob):

    def mapper(self, _, line):
        for word in line.split():
            yield (word, 1)

    def reducer(self, key, values):
        yield (key, sum(values))
```

```
# Query for gdp
def get_gdp_per_year(year):
    selector = {
        'selector': {'year': year},
        'fields' : ['gdp_value', 'iso_country'],
        'limit' : 500
    }
    result = db_gdp.find(selector)


    data = [{'iso_country': row.get('iso_country'), 'gdp_value': row.get('gdp_value')} for row in result]
    return pd.DataFrame(data)
```

Analyse data

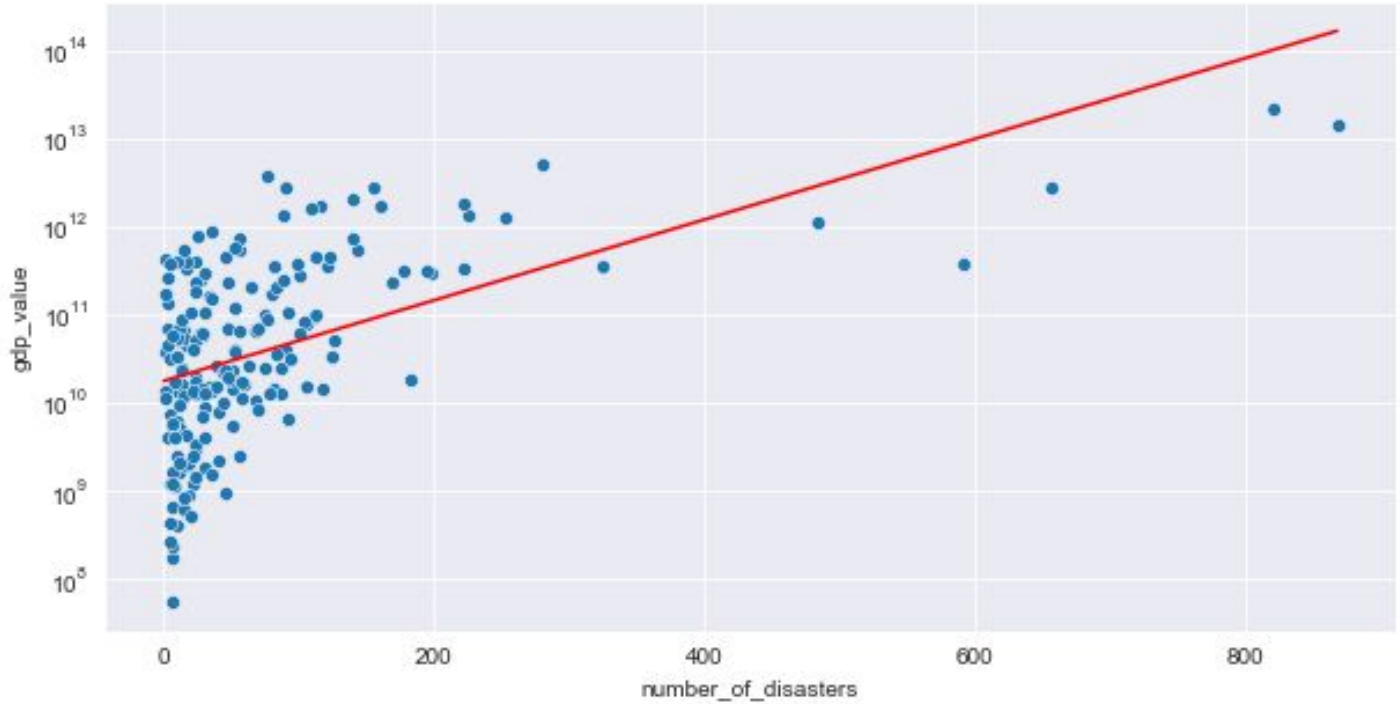


Analyse data



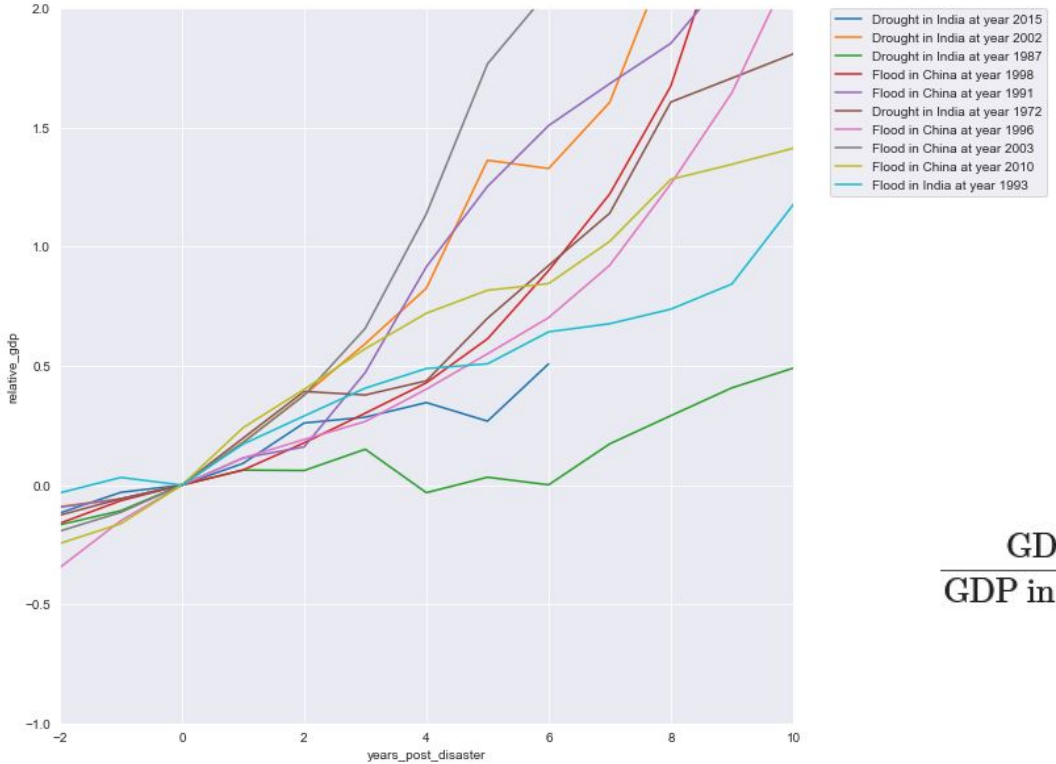
Analyse Data

Correlation: +**0.75**



Analyse data

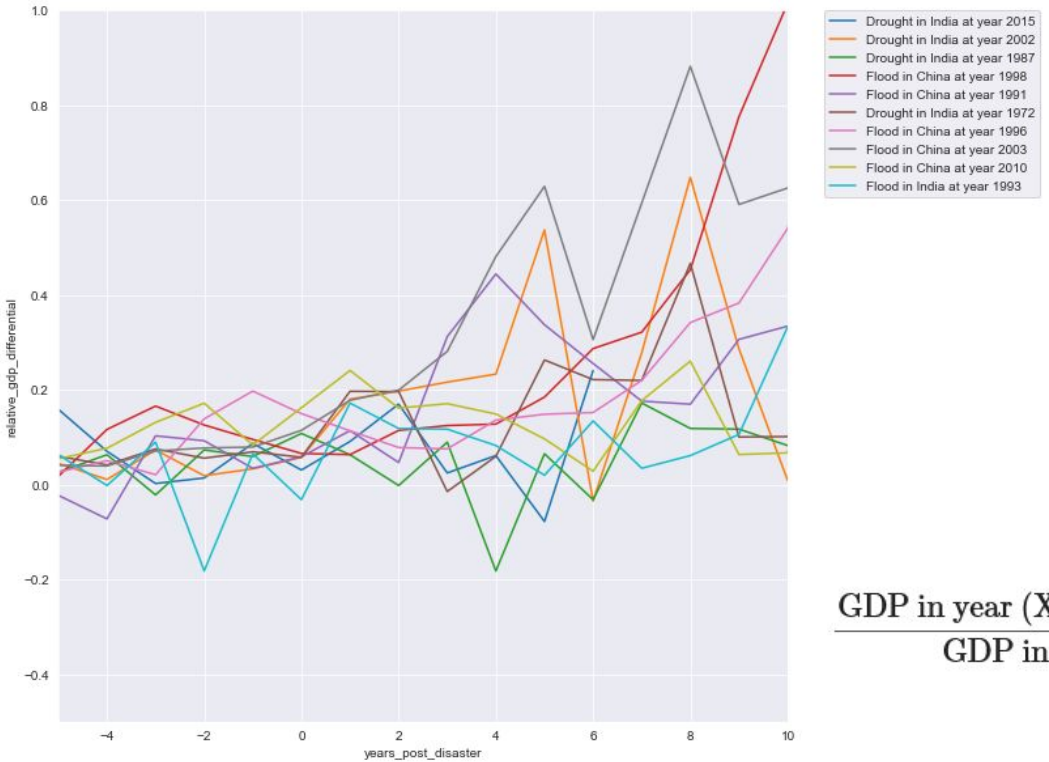
years post disaster
vs
GDP



$$\frac{\text{GDP in year X}}{\text{GDP in year of disaster}} - 1$$

Analyse data

years post disaster
vs
change GDP



$$\frac{\text{GDP in year (X + 1)} - \text{GDP in year X}}{\text{GDP in year of disaster}} - 1$$

Machine learning:

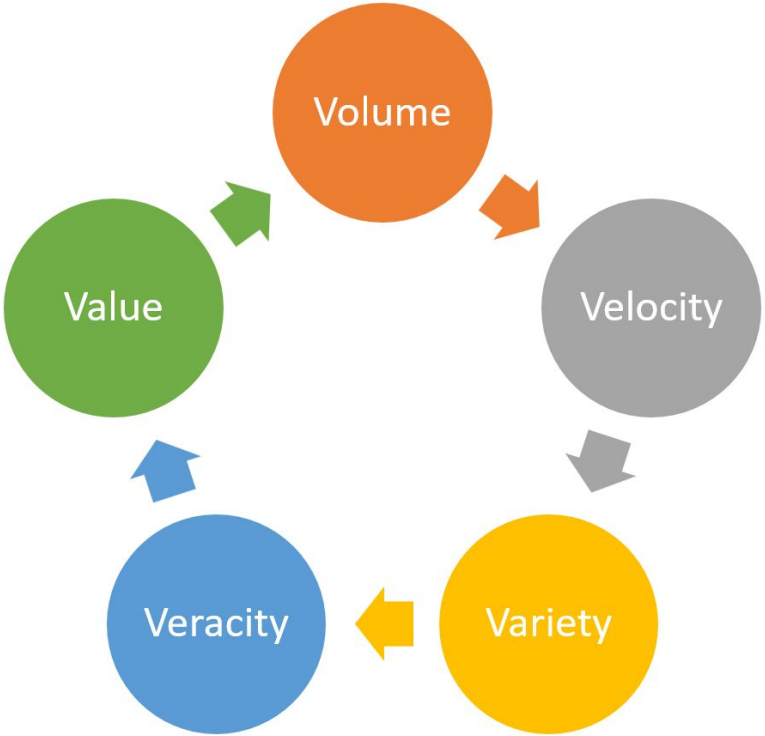
- create a connection to the CouchDB
- to take the Natural Disaster and the GDPs Data out.
- combine the data into a table.
- Dropping rows with NaN Values in the GDPs columns.
- making one-hot encoding for the ‘Type’ and ’Group’ columns.
- Split Data. The Target is gdp_change.
- Implement Linear Regression
- Calculate the Mean absolute error 18.87%

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |x_i - x|$$

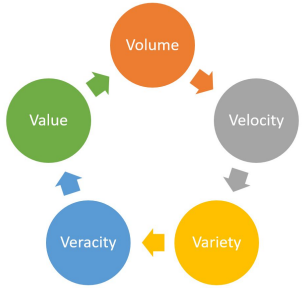
Conclusions:

- positive correlation between GDP and number of disasters: 0.75
- hypothesis is contradicted
- disasters might be also an opportunity for economic growth through reconstruction
- other/more economic indicators needed: GPD per capita, GINI
- Linear Regression Algorithm MAE: 18.87%
- Natural disasters alone are not good enough as a predictor for GDP
- Correlation does not always mean causation

5Vs



4 Levels of Data Handling



Data Source	Natural Disasters dataset: csv GDP: API
Data Storage	NoSQL CouchDB
Data Analysis	Querying: MapReduce, self-made functions Statistical Analysis: scipy and numpy libraries
Data Output	Graphic representation with seaborn and matplotlib libraries

Questions
Feedback